THE SHUFORD-MASSENGILL CORPORATION

A REPORT ON
THE EFFECT OF "DEGREE OF CONFIDENCE" IN STUDENT TESTING

H. Edward Massengill and Emir H. Shuford, Jr.

BLANK PAGE

A REPORT ON THE EFFECT OF "DEGREE OF CONFIDENCE" IN STUDENT TESTING

H. Edward Massengill and Emir H. Shuford, Jr.

Classroom testing at Muzzey Junior High in Lexington, Massachusetts is taking on a new complexion. A look at the tests being given does not reveal anything new. The test questions are of the same type and content as always. The difference is in the way the students respond to the questions. Each student responds to each question by giving his degree of confidence that each answer to the question is correct.

The students at Muzzey try to reflect their degrees of confidence as accurately as possible because it has been demonstrated to them that they can expect to make their best test score if and only if they are honest. Thus, a student who has no idea what the answer is does not try to fake or to beat the system but simply indicates his lack of knowledge.

Teachers are becoming more conscious of each student's specific strengths and weaknesses since they now know each student's state of knowledge for each question. Students are learning to explicitly evaluate information thanks to the structure supplied by Valid Confidence Testing which enables and encourages this evaluation.

THE NEED FOR DEGREE OF CONFIDENCE INFORMATION

Throughout the history of objective testing, students have responded to test questions with choices. They have used their information about an item to choose an answer. But the nature of the information leading to this choice has been lost to the teacher. And furthermore, the choice method of testing in which a student is either right or wrong has consistently reinforced the idea among both students and teachers that you either know something or don't know it.

But a little self-examination shows that knowledge is really a matter of degree. Of course, we all experience situations in which we are completely certain of some fact and situations in which we are completely uncertain. But we also experience situations in which we are almost certain, others in which we have just a fair amount of certainty, and still others where we have only a slight degree of certainty. And, of course, there are times when we experience perhaps the most embarassing situation of all: the one in which we are absolutely certain of something only to find out later, to our amazement, that the opposite is true.

The failure of current testing to preserve degree of confidence information works a hardship on both the test-taker and the test-user. The test-taker cannot indicate, in many cases, what he knows. For example, on a short-answer question the most likely answer a student can think of may be one which he believes has only a slight chance of being correct. He should, of course, write down this answer but under current testing he cannot show his lack of confidence in the answer. If his answer turns out to be incorrect, he gets the same item score as the student who put down an incorrect answer but was convinced that it was correct.

When such situations occur in a test, the effect does not "average out" but tends to be cumulative. This results in a test score which is not a valid indication of how much a student knows. In many cases the difference is large enough to change a student's letter-grade significantly. Sometimes it results in a student failing the test when he should have passed. Certainly such a result is not fair to students.

The fact that the student cannot indicate more exactly what he knows about each question hinders the teacher in diagnosing the student's specific strengths and weaknesses. The teacher never knows whether a correct answer indicates complete knowledge, partial knowledge, or a lucky guess. Or whether an incorrect answer indicates no information or some degree of misinformation.

OBTAINING VALID DEGREE OF CONFIDENCE RESPONSES

How can the testing situation be structured so that, without abandoning objective tests, we can elicit from a student his degree of knowledge? Recent developments in test theory have made possible for the first time an objective testing procedure which is capable of yielding valid information concerning a student's degree of knowledge on a test question. The new procedure is called Valid Confidence Testing. In Valid Confidence Testing, a student gives his degree of confidence that each possible answer to a test question is correct.

What assurance do we have that a student will give meaningful responses? First, the student's responses are scored in such a way that it is in his best interest to be honest in responding. Second, an individual response aid called the SCORULE™ has been developed which embodies the basic concepts of Valid Confidence Testing and aids the student in developing his response to each question. Third, there is an empirical procedure, discussed in the accompanying handouts, which can be used to determine the validity of the total system.

Thus, a workable procedure exists for obtaining the information which is essential in grading students fairly, in enabling students to explicitly evaluate information, and in providing teachers with the information they need to diagnose and treat the specific strengths and weaknesses of students, both in knowledge and in information evaluation.

TRAINING PROCEDURES FOR VALID CONFIDENCE TESTING

But can students give these responses? What kind of training is necessary? Our work at Muzzey Junior High indicates that students at the seventh and eighth grade levels can learn to give meaningful responses in the space of an hour. The higher ability students at this level pick up the procedure in less than an hour while lower ability students may require more time. Students in high school and college should certainly require no more than one hour.

A typical training session in Valid Confidence Testing begins with a question and answer approach relating degree of knowledge to test questions and showing how the Scorule is used to indicate various degrees of certainty. After a few sample questions, the students are given a practice test containing 10 to 15 items taken from the subject matter being taught in the class.

There are usually three levels of item difficulty: very easy questions, moderately difficult questions and extremely difficult questions. This helps to guarantee that the students will have a chance to respond in situations in which they have varying degrees of confidence. The practice test is given one item at a time so that individual problems can be identified and rectified early in the training session.

Besides learning to respond in the new way, the students also learn to score and classify their responses. Thus, the student sees immediately after taking a test exactly what states of knowledge led to his particular score. The classification procedure is especially important since it provides a one-digit summary of a student's state of knowledge for each question. In Valid Confidence Testing, a student is not right or wrong for a given question as he is in choice testing but rather is classified according to his degree of knowledge on the question:

If he has a very high degree of confidence in the correct answer, he is classified as well informed.

If he has a rather high degree of confidence in the correct answer, he is classified as moderately informed.
If he can eliminate some but not all the incorrect answers and is equally uncertain among the others, he is classified as partially informed.
If he has equal confidence in all possible answers, he is classified as uninformed.
If he has a high degree of confidence in an incorrect answer, he is classified as misinformed.

BASIC RESULTS OF VALID CONFIDENCE TESTING

What are the basic results to be expected from the use of Valid Confidence Testing? We will mention three major results.

First, a student's test score is more reliable and valid than the comparable choice test score for the same test.

In Valid Confidence Testing, a student's score is a function of how much confidence he puts in the correct answer to each question. This function must be a special non-linear function in order to make it in the best interest of the student to be honest in responding.

At first thought, it might seem that the use of Valid Confidence Testing would introduce additional unreliability into the total test score. After all, though we can fairly consistently identify situations in which we are completely sure or completely unsure, they may be some instability in making a judgment of degree in situations in which we have moderate certainty. But a close analysis shows that such situations contribute very little to instability as compared to the instability that is eliminated by Valid Confidence Testing.

Valid Confidence Testing eliminates the major source of instability which arises when one is uninformed between the correct and one or more of the incorrect answers. This type of situation is commonly referred to as a guessing situation. The presence of one of these guessing situations in a test, for a given student, can overwhelm the presence of many of the "moderate certainty" situations.

Additionally, the more experience a student has taking tests as Valid Confidence Tests, the more reliable his confidence judgments should be. In fact, perfect reliability of score can be approached in Valid Confidence Testing whereas no amount of experience with test-taking can promise this result in choice testing. Thus, the use of the Valid Confidence Testing will almost invariably result in a more reliable score.

In terms of validity, if a student has complete confidence in some answers for each question, his choice score will be as valid as his confidence score on that test. But, as a student has situations in which he has less than complete confidence, his confidence score will be much more valid.

To see how this happens, let us look at a short-answer test. Suppose that two students both give the same number of correct answers but one student is completely sure of all of his answers while the other is uncertain for some of them. Under current testing procedures, both students would receive the same score. Thus no distinction is made between these two students, one of whom has more knowledge than the other. In Valid Confidence Testing, the student with more knowledge would receive a higher score.

Suppose that on the same test two students have the same number of correct and incorrect answers. But one student has high confidence in all of his correct answers and low confidence in his incorrect answers while the other student has high confidence in all of his answers. Again under current testing procedures both students would receive the same score.

But the student who can distinguish between what he knows and what he doesn't know has more knowledge than the student who can't. A Valid Confidence Testing score would make this distinction.

One result of this failure of current testing procedures to make distinctions such as the ones discussed above is that only the very best and very worst students in a class receive a valid score while most of the others obtain scores which grossly underestimate their degree of knowledge. For example, we have had some students improve their scores by as much as 20 points on a percentage scale by taking the test as a Valid Confidence Test.

The second basic result of using Valid Confidence Testing is that the specific strengths and weaknesses of a student are clearly identified.

It is impossible to determine from a student's choice alone on a question, whether he had much, moderate, or little confidence in the answer he chose. For example, he may have had no more confidence in the answer he chose than in any of the other answers.

But it is just this information a teacher needs in order to determine what instructional steps the student needs next. Valid Confidence Testing gives this much-needed information in the form of a student-by-item table.

Using the table, the instructor can look at a given student's classification pattern and determine where he needs help and what kind of help he needs. He can look at the individual items and determine which ones he has successfully taught and which he has failed to get across. He can form groups of students who need help in the same area or areas. He can assign supplementary work to students who know the material.

The third basic result of the use of Valid Confidence Testing concerns the explicit evaluation of information by students taking their tests this way.

Our empirical work to date in Valid Confidence Testing indicates that students have varying degrees of confidence for various test questions, that they can learn rather easily to use the Scorule to indicate these degrees of confidence, and that they honestly indicate their degrees of confidence. But these same results indicate that many students are not very good at evaluating information.

For example, many math students believe that complete confidence is justified merely because they work a problem and arrive at an answer. The result is many instances of misinformation for a student on a test. These students are not being dishonest. They really feel this way. But these students need help, not only in increasing their knowledge about the subject matter but in learning to evaluate the knowledge that they have.

Valid Confidence Testing results graphically point up such problems not only to the teacher but also to the student. And the logic behind Valid Confidence Testing suggests that merely taking tests as Valid Confidence Tests can improve a student's ability to evaluate information. Certainly, explicit training through the use of Valid Confidence Testing materials can lead to such improvement.

THE USE OF VALID CONFIDENCE TESTING IN CAI

How would Valid Confidence Testing work in a CAI program? What additional hardware and software would be needed?

We believe that the basic need in adapting Valid Confidence Testing to CAI is not in hardware but

in software. In our earliest work in this area we used the computer as a response aid. The computer was programmed so that the student could indicate his degree of confidence using a light pen on a scope. After more than two years of studying the problem, we believe that there is a better approach both from the standpoint of economics and from ease of use.

We can see, for example, a student using a Scorule to obtain his confidence response and entering it into the computer through a scope or keyboard. The computer would then analyze the response and make a decision as to the next step. Using this approach the basic need is for a sub-routine to accept the response pattern, calculate the item scores and the total score, and classify each response pattern into one of the mutually exclusive and exhaustive categories mentioned above. The rules for these operations are already well-defined so that it is just a matter of reducing them to computer language. Of course, as we shall see, there are various levels of complexity at which Valid Confidence Testing can be used in CAI and the use of the more complex levels would require further programming.

Let us briefly look at three possible applications of Valid Confidence Testing to Computer Assisted Instruction. First, since Valid Confidence Testing can be used with any objective test, it seems natural to consider substituting this new response technique for unit tests, both pre-tests and post-tests. For existing programs, Valid Confidence Testing can be introduced without modifying the questions. The test scores, per se, would be more reliable and valid. Students would have had a chance to explicitly evaluate their knowledge and would have a better understanding of their own strengths and weaknesses. Information would be available concerning where a student needs additional work.

A second application would be to use Valid Confidence Testing for making branching decisions within a computerized course of instruction. Branching decisions based on degree of confidence rather than choice would greatly reduce errors in branching students to appropriate instructional sequences.

The branching routine could be as simple or as complex as desired. For example, in a program in which students are now branched according to whether the student is correct or incorrect, the decision could be changed to whether the student knows or doesn't know the answer. In other words, some degree of confidence cut-off point, such as 90% confidence in the correct answer, could be used in deciding whether or not the student knows or doesn't know the concept.

Stepping up another level in complexity, suppose the student is currently being branched on the basis of which answer he chooses. Here the student presumably choses an answer even when he is completely uncertain. Thus no matter which section he is sent to, the instructional sequence will be only partially adequate, if at all. Valid Confidence Testing could be used to branch the student to one of five categories for a four-answer question:

Category 1: the student knows the answer.
Categories 2, 3, and 4: he is misinformed on a particular incorrect answer.
Category 5: he doesn't know the answer.

A slightly more complex branching decision would be to add categories which specify that the student be branched to a particular sequence if he rules out a particular incorrect answer or answers but is uncertain among or between the rest. There could also be additional misinformed categories for cases in which the student is uncertain between two or more incorrect answers.

And, of course, the decision can be complicated even more by including possibilities such as a

sequence for the moderately informed student.

The main point is that the branching decision can be as simple or as complex as the situation demands. But even in the simplest case, the potential benefits are great not only in terms of fewer branching errors but also in terms of enabling the student to more explicitly and meaningfully evaluate his knowledge. And the availability of simple branching decisions means that Valid Confidence Testing can be incorporated now into existing programs with minimal disruption.

And finally there is a third application of Valid Confidence Testing to CAI which promises extremely efficient instruction. This application involves a type of sequential testing in which questions concerning a topic are ordered logically and/or empirically in such a way that if a student knows the answer to a question at a given level, he most likely knows the answers to all questions below this level.

We will mention three implications of this approach, which we characterize as Tutorial Testing. First, students can be more efficiently tested. A student doesn't have to attempt questions which we can be almost certain that he already knows. Second, when a student is having trouble very efficient probing can be done in an attempt to locate the specific source of the problem. For example, when a student doesn't know a question at a given level, he can be sent down to the next level where there might be several questions each with its own branches. Third, the student can be assigned appropriate instructional sequences, which may or may not be on-line.

IN SUMMARY

1. It is both possible and feasible to introduce Valid Confidence Testing into CAI. Students have shown that they can use the approach in much more difficult non-computer situations.

2. We believe that the use of Valid Confidence Testing can aid users of CAI in further dispelling the notion that computers dehumanize education. Since the information obtained is more like what would be available if a teacher observed each student closely as he took the test or administered an individual test to the student, the computer could be much more responsive to the needs of the student.

DEGREE-OF-CONFIDENCE RESPONSES FROM VALID CONFIDENCE TESTING

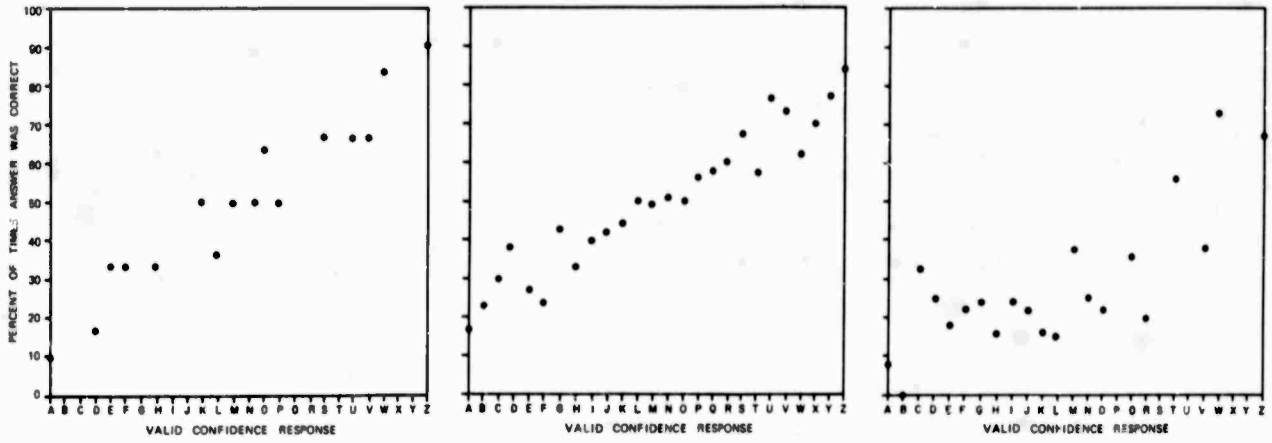
Our analyses of the Muzzey Junior High School and other data show that responses to a Valid Confidence Test are much more reliable than those choices made in the old multiple-choice and fill-in-the-blank tests. Remember though that, however high the reliability, the responses still could be meaningless and totally without validity.

Let us take a simple and naive view of validity. A Valid Confidence response should reflect the actual chance that an answer is in fact the correct one. If a student's responses predict his actual chances of success in applying his knowledge, then the data from a Valid Confidence Test can be interpreted in the most direct fashion possible.

Two 20-item Short-Answer Tests In Junior High Mathematics
Same 25 Students for each Test
Third and Fourth Valid Confidence Tests

One 17-item Short-Answer Test In Junior High Science
52 Students in Two Classes
Third Valid Confidence Test

One 12-item Multiple-Choice (5) Test In Junior High Science
49 of the 52 Students to left
Fourth Valid Confidence Test



The horizontal axis of each graph above represents degree of confidence where A = .00, B = .04, . . . , Z = 1.00. Each point on a graph was obtained in the following way. First, the number of times all students used a particular degree of confidence was counted. Then the percentage of times it was used on a correct answer was computed.

It should be noted that the left graph represents two tests, the third and fourth Valid Confidence Tests, taken by one class. The middle graph represents the third Valid Confidence Test taken by two classes of students. The right graph is based on these same two classes who were taking their fourth Valid Confidence Test.

The danger here is that this validity can show itself only if the students understand the SCORULE response aid, respond honestly, evaluate information well, and do not have much misinformation. It is a stringent test of validity.

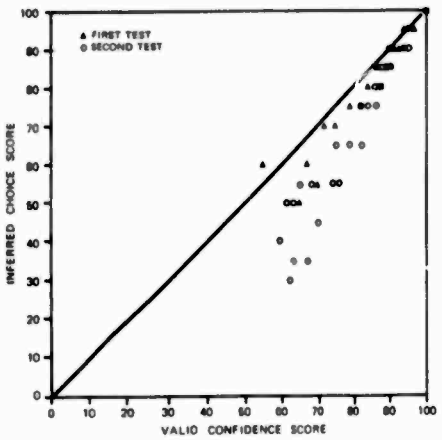
The data shown in these graphs certainly pass anyone's "eyeball test" for the existence of a direct relation between response and chance of success. There remains little room for doubt that Valid Confidence responses can be realistic predictors of success in the real world.

A STANDARD OF PERFORMANCE FOR THE OLDER TESTING METHODS

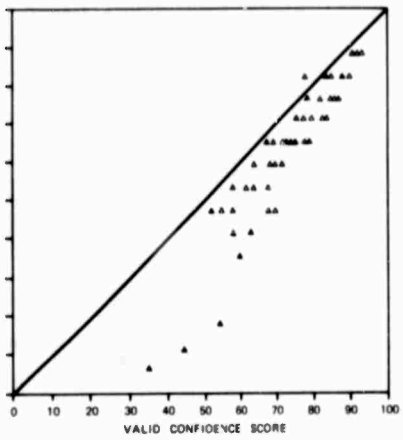
TOTAL TEST SCORE

Valid Confidence responses are more reliable and valid than the choices made in the old multiple-choice and fill-in-the-blank tests. There are logical and philosophical arguments which indicate that the scoring system of Valid Confidence Testing is the right way to value knowledge. It seems natural, therefore, to consider what might be lost when an objective or semi-objective test is administered as a choice test. In particular, let us look at the total test scores yielded by the two methods. Total test score is important because it is the basis for course grades in schools and for personnel decisions in schools, industry, and government.

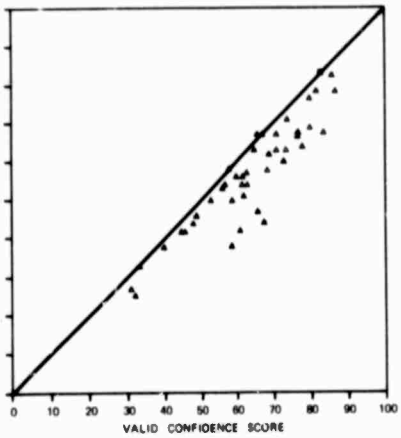
Two 20-item Short-Answer Tests



One 17-item Short-Answer Test



One 12-item Multiple-Choice (5)



The three graphs in this section are based on the same groups as those shown in the previous section. A blackened triangle or square represents two students while an open triangle or square represents one.

The choice test scores are inferred from the Valid Confidence Testing data. When a short-answer test is given as a Valid Confidence Test, the student picks the most likely answer for each question and gives his degree of confidence that that answer is correct. Thus, the choice score for short-answer items can be obtained by dividing the number of right answers by the number of questions.

For multiple-choice items, we assume that a student would have chosen the answer in which he had maximum confidence. There are many cases when the student had maximum confidence both in the correct answer and one or more incorrect answers. Here the choice score for an item is the student's expected score. For example, if he were completely uncertain between the correct answer and an incorrect answer, his expected item score would be .5. If he divided his confidence so that one-third was on the correct answer and one-third on each of two incorrect answers, his item score would be one-third.

If the paired scores (one for the Valid Confidence Test; one for the choice test) for a class of students fall exactly on a straight line, then the same grades (or the same personnel decisions) would be made if choice testing were resorted to. The graphs shown here indicate that this is not the case. Choice testing does not even yield the same rank ordering of students. Thus, the use of choice testing means that many students are graded unfairly or that many personnel decisions would be wrong.

Notice further that the choice test scores tend to be too low, especially for the poorer students. Thus, choice testing underestimates the achievement of many students.

A STANDARD OF PERFORMANCE FOR THE OLDER TESTING METHODS

ACCURACY OF DIAGNOSIS

Valid Confidence responses are more valid and reliable than the choices made in the old multiple-choice and fill-in-the-blank tests. It seems natural, therefore, to consider what might be lost when an objective or semi-objective test is administered as a choice test. In this instance, let us look at the accuracy of choice testing in diagnosing the student's state of knowledge. Accurate diagnosis helps in understanding the student, in evaluating instruction and item writing, and in guiding instruction.

	TWO 20-ITEM SHORT-ANSWER TESTS				17-ITEM SHORT-ANSWER		12-ITEM MULTIPLE-CHOICE	
	STRICT	WEAK	STRICT	WEAK	STRICT	WEAK	STRICT	WEAK
PERCENT OF STUDENTS	88	75	92	92	100	96	96	84
PERCENT OF ITEMS	19	15	38	32	41	30	37	17

The first row of the table shows the percent of students who would have been incorrectly diagnosed if the test had been given as a choice test.
The second row shows the average percent of items for which each student would have been incorrectly diagnosed.

In Valid Confidence Testing, five mutually exclusive and exhaustive categories have been defined to provide the teacher with a one-digit summary of each student's state of knowledge for each question.

W, well-informed, represents a high degree of confidence in the correct answer.
I, moderately informed, represents a fairly high degree of confidence in the correct answer.
U, uninformed, represents equal confidence in all answers.
P, partially informed, represents high confidence in the correct answer and the same confidence in one or more of the incorrect answers.
M, misinformed, represents low confidence in the correct answer and, thus, high confidence in one or more of the incorrect answers.

If we relate "correct" in choice testing to "W", then logically we must relate "incorrect" to "M". This implies that anytime a student is classified as U, I, or P for a Valid Confidence Test question, he would have been misdiagnosed if the test had been given as a choice test.

A more stringent criterion is to say that the choice test makes an error whenever a student has anything other than complete confidence in any of the answers.

From the error rates shown in the table above, we must conclude that teachers are getting a distorted view of most students and a significant percentage of the items.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)

The Shuford-Massengill Corporation
Box 26, Lexington, Massachusetts 02173

2a. REPORT SECURITY CLASSIFICATION: UNCLASSIFIED

2b. GROUP

3. REPORT TITLE

A REPORT ON THE EFFECT OF "DEGREE OF CONFIDENCE" IN STUDENT TESTING

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Scientific Interim

5. AUTHOR(S) (First name, middle initial, last name)

H. Edward Massengill & Emir H. Shuford, Jr.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
February 1968	9	0

8a. CONTRACT OR GRANT NO	9a. ORIGINATOR'S REPORT NUMBER(S)
AF 49(638)-1744	
b. PROJECT NO: 920F - 9719	SMC R-11
c. 6154501R	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d. 681313	AFOSR 68-2161

10. DISTRIBUTION STATEMENT

1. This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
TECH, OTHER	Air Force Office of Scientific Research 1400 Wilson Boulevard (SRLB) Arlington, Virginia 22209

13. ABSTRACT

The use of Valid Confidence Testing at Muzzey Junior High School in Lexington, Massachusetts has demonstrated that students at all ability levels can learn to use Valid Confidence Testing materials, that they are honest in responding, that they have varying degrees of confidence for test questions, and that the responses are valid.

It has also been found that the average ability students can learn to give these valid responses in a one-hour training session and can learn to score and interpret their responses during a second one-hour session. Once they have been trained, they can take a regular classroom test, score the test and interpret their states of knowledge during a one-hour class period.

The basic results of the use of Valid Confidence Testing at Muzzey indicate first that scores obtained from classroom tests are more valid than they would have been if the tests had been administered as choice tests. Second, that through Valid Confidence Testing the specific strengths and weaknesses of a student can be clearly identified in a way not possible in choice testing. And, third, that the explicit evaluation of information by students gives them additional insight into their knowledge.

Valid Confidence Testing works in the classroom and it can be expected to work in Computer-Assisted Instruction both in unit tests, with branching instructional programs, and in the largely unexplored area of sequential testing.

DD FORM 1 NOV 65 1473

UNCLASSIFIED

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Valid Confidence Testing						
Test validity						
Computer Assisted Instruction						
Training						